DOCUMENT RESUME

ED 273 088 FL 015 805

AUTHOR Madsen, Harold S.
TITLE Dimensions of Computerized Rasch Applications in ESL Testing.
PUB DATE [86]
NOTE 45p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; Computer Software; *English (Second Language); Item Analysis; Item Banks; *Language Tests; Microcomputers; Scoring; *Statistical Analysis; *Test Construction; *Test Items
IDENTIFIERS *Rasch Model

ABSTRACT

The most appropriate statistical model for the small-scale (n<100) studies common in language testing research is the Rasch one-parameter logistic model. The Rasch model provides a wide range of options for conducting research, refining existing examinations, and developing tailored (computerized adaptive) language tests. Three investigations using this analysis in testing for English as a second language (ESL) illustrate its usefulness. One looked for possible item bias in an ESL placement test battery, a second investigated the identification of inappropriate answers suggesting cheating, and a third involved the development of a computerized adaptive grammar and reading test. Several charts and a three-page list of references conclude the report. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED273088

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Harold Madsen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

DIMENSIONS OF COMPUTERIZED RASCH APPLICATIONS IN ESL TESTING

Harold S. Madsen

This paper is intended to demonstrate the range of options that Rasch analysis makes available to those in language testing for conducting research, refining existing examinations, and developing tailored (computerized adaptive) language tests.

Contemporary Item Response Theory research has resulted in the generation of various statistical models designed to facilitate Latent Trait analysis. But the one most appropriate for the small-scale studies so frequently encountered in language testing investigations (with an N of 100 or less) is the Rasch one-parameter logistic model (Lord 1983, Rasch 1966). Not without its critics (Goldstein 1979), Rasch analysis has been demonstrated to be robust in a variety of areas including unidimensionality requirements (see Henning, Hudson, Turner 1985).

Based on the relationship between item difficulty and student ability, Rasch analysis has been utilized to enhance the precision of item calibration (Englehard 1980), conduct sample-free item analysis (Wright and Masters 1982), facilitate development of alternate test forms (Alfred 1986), develop item banks (Rubin and Mott 1983, Woods and Baker 1985), and link tests (Wright and Stone 1979). The procedure has also been used to calibrate speaking test interview items allowing for partially-

FL015805

correct responses, and even to analyze a scalar self-rating language ability instrument and a writing scale (Davidson and Henning 1985, Henning and Davidson 1986).

The versatility of the Rasch procedure will be indicated by drawing upon three ESL research and development efforts utilizing Rasch analysis.

Item Bias Studies.

The first study investigated possible item bias in an ESL placement test battery. Concern over test bias has often been expressed both within the profession (Briere 1973, Oller and Perkins 1978, Farhady 1979, Mohan 1979, Madsen 1982) and without (Nakano 1977, Houts 1979). Bias can be generated in a variety of ways, ranging from how a test is administered, to the inclusion of items with built-in cultural assumptions unrelated to performance objectives.

Item Response Theory is particularly well suited to examine aberrant item performance; and it is also capable of identifying those students for whom the test is inappropriate (Hulin and others 1983, 152-184). Procedures for carrying out Rasch analysis have ranged from hand calculation to the use of mainframe computers using routines such as BICAL (Wright and Mead 1976). For our bias investigation, however, we utilized Rasch software recently developed by Benjamin Wright and Michael Linacre for the IBM personal computer: Produced by Mediax Technologies, Inc. under the label Microscale Plus, this microcomputer program utilizes a Supercalc spreadsheet with a

grid 60 x 254: each axis able to handle either students or items. (A very recent update for the microcomputer utilizes the Systat spreadsheet: While this handles vast numbers of items and subjects, its graphics are decidedly inferior to those generated by Supercalc.) The Microscale program accommodates well more than 50 percent missing data.

Data is entered onto the spreadsheet (1 = correct, 0 = wrong, and missing data are left blank). Two convergence routines, PROX and UCON, are run. A quick and general fit of items and persons is effected through PROX, which transforms data to a logit metric. This is followed by a far more complex and more accurate "general unconditional" method called UCON, which "estimates the person parameters simultaneously with the item parameters" (Wright and Masters 1982, p. 61). These result in calculations on the same "logit" scale of person ability and item difficulty.

> A person's ability in logits is their natural log odds for succeeding on items of the kind chosen to define the "zero" point [or mean] on the scale. And an item's difficulty in logits is its natural log odds for eliciting failure from persons with "zero" [or average] ability. (Wright and Stone 1979, p. 17)

Figure 1 illustrates a logit scale, the upper bar graphs representing students, the lower representing test items; weaker students and easier items are at the left.

In addition to ability-difficulty calibration in logits, our

analysis also calculates "fit" statistics, indicating statistically and graphically how well each item measures those taking the test. The "information weighted" statistic or infit "focuses on the overall performance of an item or person" in the range with the closest matchup of item difficulty and person ability. The "outlier-sensitive" fit statistic or outfit identifies "unexpected events"--student or item outliers, which are generally "far removed from one another on the variable" (Wright and Linacre 1985, G-11). For example, the high positive outfit of a high-ability person could be the result of missing a few easy items due to carelessness. See Figure 2.

Contrived Bias Checks. Since it would be unethical to administer an exam that was intentionally biased, it was arranged to present a Spanish cognate vocabulary exercise in an English Language Center reading lab, where results would not contribute to student grades but could provide practice in vocabulary building (Madsen and Larson 1986). The 60 ESL subjects included 39 who were expected to perform well on the exercise because of their language background (36 native Spanish speakers, 2 Portuguese, and 1 Quechua with good proficiency in Spanish), and 21 without special advantage as far as language background is concerned (10 Japanese, 4 Chinese, 2 Arabic, 1 Farsi, 1 Korean, 1 Vietnamese, 1 Thai, 1 Greek). The 60 subjects had been placed at 6 levels of instruction from beginning to advanced--Spanish speakers predominating in number at each level.

The 30-item exercise consisted of a single-word stem

followed by 4 options. To accommodate beginning-level subjects (consisting only of Spanish-speaking and Japanese-speaking students), the test was presented in versions utilizing Spanish or Japanese instructions. The mean for the Spanish-Portuguese group was 26.4; the mean for the remainder was 17. The mean for the combined group was 23.1 with a standard deviation of 5.7 and a reliability of .86.

The bias was identified, initially in the map (Figure 3) and then in the outfit display (Figure 4). The map predictably generated a bimodal distribution of students, the solid portion of the bars representing students with Spanish or Portuguese background. In the outfit scattergram, 8 out of 11 outliers (73 percent) were non-Spanish speakers, although this group constituted just a third of the subjects evaluated. While not all non-Spanish speakers were identified, the vast majority of those with aberrant responses were students against whom the test was biased.

A second contrived bias check consisted of our taking an easy item (number 11) from a 60-item multiple-choice grammar test that had been administered to over 150 ESL students and causing it to "behave abnormally." (A significant amount of latent trait research consists of computer simulations [Weis 1983].) By modifying the data, we made it appear that five students from the top quartile had missed this easy question. The Rasch analysis now identified item 11 as an outlier: Consulting Figure 5, note the dramatic vertical shift of this question when it is "missed"

by even a few advanced students. In a real-life situation, one would survey those unlikely to have missed the item to discover the source of bias, such as a small group of arts students taking an ESP test intended for science majors.

Checking for Actual Bias. It was hypothesized that Rasch analysis of an actual ESL test administration would identify bias in the form of identifiable patterns of outliers or poor fit statistics. One such pattern, it was felt, would be associated with student language background. Other potential patterns of poor fit included academic major and ability level of subjects.

A total of 183 ESL students ranging from beginning to intermediate proficiency were administered a five-part ESL placement battery. The three multiple-choice sections of this battery (grammar, reading, and listening) were Rasch analyzed to identify possible bias. Analysis of variance was also utilized to confirm observed differences.

Of the data generated, the outfit statistic was focussed on inasmuch as this is most sensitive to aberrant responses. Outfit data were likewise used in ANOVA calculations, with language background as the independent variable, level as moderator variable, and outfit data as the dependent variable.

Outfit plottings of the Rasch data failed to confirm the initial hypothesis; and ANOVA calculations on outfit data by language background likewise failed to indicate statistically significant differences. In short, the ESL battery appeared not to be biased in favor of any particular language group.

We next looked for patterns of outliers in areas other than language as we examined the separate Rasch analyses of the three subtests. While outliers on the reading test were rather evenly distributed, discernible patterns emerged on the listening and grammar tests. Outliers emerged almost exclusively from the lower half of the students on the listening test and from the lowest third of the grammar test (Figures 6 and 7). The ANOVA calculation, with proficiency level as independent variable and outfit as dependent corroborated the Microscale displays of outliers: Statistically significant fit differences occur at varying ability levels (Table 1).

This unusual amount of erratic behavior by low-ability students suggests that the grammar and listening subtests may be too difficult for a number of beginning-level people. And in this sense, these sections appear to be somewhat biased in favor of intermediate language learners. This finding tends to support the second hypothesis of bias (or outlier patterns) related to factors other than language--here, ability level.

Cheating Studies.

Equally important to bias in IRT applications is the broad area of appropriateness measurement, which has been defined as "any method of detecting individuals for whom test scores fail as measures of the latent" trait, ability, or interest (Hulin and others 1983, 111). In short, this area of investigation examines test taking anomalies.

These aberrations can be classified into two categories:

unintentional behaviors and deliberate tampering with results.

Wright and Stone (1979, 165-180), Levine and Drasgow (1982), Hulin and others (1983, 112-114) mention several examples of unintentional behaviors that can provide misleading test results. These include a student's guessing, being inattentive or bored, "fumbling" because of confusion over the intent of test questions or exam directions, stumbling on items because of their cultural bias, not fully completing an exam, utilizing suboptimal omitting strategies, or simply looking at items (as on an IQ instrument) from a novel or creative point of view. Aberrant results can also stem from special handicaps such as limited language proficiency when taking a test designed for native speakers of the language or hearing deficiencies when taking an aural comprehension test. Misleading results can also stem from an atypical educational background, from mismarking a machine-scoreable answer sheet, or being "very conservative" in one's use of "partial information" (Levine and Drasgow 1982).

Deliberate tampering with results takes many forms. There have been occasions when teachers have tampered with answer sheets to improve student scores. There are occasions when answer keys are obtained and sold to some candidates in advance of the test. More common is an examinee's copying from another examinee during a test administration. Normally only a block or two of questions is copied or memorized. Responses on attitude measures are at times modified by the respondent to match perceived expectations on the part of those using the instrument.

And examinees have been known to suppress scores on a minimal proficiency exam battery such as those administered by a government for military draftees.

Since conventional statistical analyses of tests have had limited success in identifying inappropriate measures obtained on tests (aberrant scoring having little or no effect on correlations, for example), Hulin and others (1983) survey research on various IRT-based models for detecting inappropriate responses on tests, including the one-parameter model discussed in this paper. They conclude that Rasch is a viable option (p. 141).

Initial TOEFL Cheaters Studies. It was decided to conduct appropriateness studies that focussed on the cheating problem, and again to utilize a Rasch-based computer program for the analysis. While the first two investigations failed to identify cheaters, they set the stage for follow-up work which did confirm the robustness of the one-parameter Rasch procedure in detecting the aberrant behavior of those who would tamper with test items.

The initial investigation involved two known cheaters on the TOEFL (Test of English as a Foreign Language). Two sets of answer sheets, with identification removed, were obtained from Educational Testing Service (at Princeton, New Jersey), the agency responsible for administering TOEFL. Selected from exam papers at five centers, there were 30 subjects that took TOEFL form 15; and there were 29 that took TOEFL form 7. One unidentified cheater was included in each group.

It was hypothesized that in order not to flag the person for

whom the cheating was being done, the cheater would attempt to suppress his score by arbitrarily missing random test items. If this happened, it was anticipated that the computer outfit analysis would identify the misfitting test taker: an able student missing a number of relatively easy items.

The two groups were Rasch analyzed separately utilizing the Microscale Plus routine, but no clearly identifiable outliers were identified (Figure 8). In short, the cheaters had not been detected. The reason for this became apparent, however, after the ID numbers of the cheaters were provided by ETS: They had performed so well (with scores of 610 and 623) that it was apparent they could not have suppressed their scores.

Therefore ETS agreed to send two additional groups of papers (identification removed as before) with cheaters whose scores were well under 600: one with a score of 557 and the other with a score of 563. One was included in a group of 44 students that took TOEFL form 9, and the second in a group of 46 students that took TOEFL form 8. As in the first TOEFL study, data from five different testing centers were utilized.

Again the outfit plots (Figure 9) generated by Rasch analysis of these second two groups failed to identify the two known cheaters. Interestingly, the analysis did, however, identify an outlier with a dramatically high outfit score; this probable cheater was reported to TOEFL officials.

But the known cheaters either had not attempted to suppress their scores, or such an effort had been so minimal that the

analysis was unable to detect it. Levine and Rubin (1979) investigated how effective a three-parameter model was in identifying item tampering in spuriously high and spuriously low test alterations. After examining 4, 10, 20, and 40 percent score modifications of artificial data, they discovered that only when tampering reached the 20 percent level or higher did they achieve good detection rates.

Cheating Studies Utilizing Artificial Data. We now attempted to determine how effective Rasch analysis would be in identifying a small percentage of manipulated scores. Drawing upon data generated earlier for 183 ESL subjects, we selected two papers at the low end of the scale and rescored them so that correct answers were given to five items in the top quartile of difficulty (see Madsen and Larson 1986, 59). Rasch analyzing the data again, we found that the outfit scores for these two papers had jumped dramatically: from approximately 1 to well over 4--the highest outliers of the entire group of 183 students (Figure 10).

In contrast with this spuriously high score situation, it was next decided to devise a more realistic setting, one parallel to that hypothesized for professional cheaters on the TOEFL. Using the cheater's paper on TOEFL form 7, we manipulated a "miss" of one in every ten responses (resulting in a score of 577). The data was again Rasch analyzed. Now with a <u>suppressed</u> score, the paper had the highest outfit of any in the group (Figure 11). But while such an outfit score could help corroborate allegations of cheating, it might not always flag decidedly

aberrant test behavior.

Like Levine and Rubin (1979), we needed to evaluate a situation well beyond the 10 percent level of manipulation.

For this, a TOEFL scenario was devised, involving a supposed native speaker cheater with a 20 percent score suppression. This contrived test paper was added as examinee number 30 in the TOEFL Form 7 group. The data were again Rasch analyzed, this time with impressive results. The native-language "cheater" had an outfit score approximately three times as great as the highest regular student (Figure 12). In short, Rasch analysis is able to detect aberrations in test responses. And while the last two studies examined the effects of suppressing test scores, the procedure would be equally effective in detecting inflated scores resulting from copying or memorizing blocks of questions.

Development of a Computerized Adaptive Test.

While computerized adaptive tests are far from being a panacea as far as language testing is concerned (Larson and Madsen 1985), they do hold out the promise of helping to meet two concerns of psychometricians--the need for precision and the desirability of invariant scaling. For a survey of the several advantages and limitations of computerized adaptive tests, see Tung (1986) and Canale (1986). For a more in depth explanation, see Section IV of Weiss (1983).

Background. Decades ago it was recognized that the closer the difficulty range of items approximated the ability level of the examinee the more precise the calibration would be of one's

mental ability. The one-on-one "tailored" testing that characterized early IQ measures can in effect be replicated with the branching routines of today's computerized adaptive tests, which select items tailored to the ability level of each individual. And Rasch-analyzed items are calibrated with a precision not previously available through conventional procedures.

Invariant scaling is likewise facilitated by IRT procedures such as Rasch analysis. Until recently, the difficulty level of test items depended on the proficiency level of the persons taking the exam. Now by means of Rasch analysis (or other latent trait procedures), logarithmic calculations determine the probability of each item's being answered correctly by persons at varying ability levels; and they likewise determine the probability of each individual's making a correct response to any given item. As we have seen (Figure 1), the microcomputer printout of the Rasch analyzed test calibrates test items along a logarithmic (logit) scale independent of examinees. The resulting calibration makes it possible to administer test items appropriate to examinees of essentially any ability level.

Development of the Computerized Adaptive Test. A computerized adaptive language test (CALT) was field tested in December 1985 at Brigham Young University. Covering the grammar and reading objectified segment of the English Language Center Placement Battery, this CALT instrument complements three other sections of the ELC placement battery--the oral interview, the writing section, and the listening comprehension section. (The

latter will be field tested in a CALT version in August 1986.)

Throughout this decade at the ELC, paper-and-pencil subtests have been administered following an oral interview. Those with low scores on the oral interview have taken a low battery exam, and those with higher scores have taken the high battery. These two batteries have placed students at five levels from beginning to advanced proficiency (500 on the TOEFL). The low and high batteries were developed with overlapping items. The top third of the low-battery grammar test items are identical to the lower one-third of the high battery items. Three forms of the test battery have been developed.

Each low- and high-battery paper-and-pencil grammar test consists of 60 four-option multiple-choice items. The low-battery reading test consists of 60 three-option multiple-choice items--each following a sentence or two of academic prose. The 45-item high battery reading test consists of sentence-level items as well as reading passage items.

Preparation of the CALT requires the development of a carefully calibrated bank of test items. This is carried out in a two-step process: first, by calibrating individual test sections, and second, by linking the calibration of each test section to the other. The process of calibration and linking will be illustrated with reference to the grammar tests.

Data from the current grammar subtests were used in developing the ESL CALT grammar version. We began by calibrating the Form 1 Low Battery Grammar Test, then Form 1 High Battery

Grammar, next Form 2 Low Battery Grammar, and finally Form 2 High Battery Grammar. By means of the overlapping items, the low and high batteries of Form 1 were linked on the same scale. Form 2 grammar tests were linked in the same manner. But to link Form 2 with Form 1, it was necessary to incorporate 20 items from Form 2 on Form 1 and administer this special test to ELC students. These 20 items, anchored in Form 1, were then used to recalibrate Form 2 so that each item on both forms was now on the same scale.

What we have described is but a brief summary of the process. After each form of the grammar test has been constructed, administered, and scored, the data are entered on the Supercalc spreadsheet of the (Rasch) Microscale software program for the IBM pc. With the data entered, the two statistical convergence routines (PROX and UCON) are applied to arrive at very precise estimates of item difficulty and student ability. Then it is possible to print hard copies of statistical results related to both students and items including, for the latter, numbers of attempts to answer the item, successful completions, logit calibrations, standard error, as well as infit and outfit statistics. Other printouts include the bar graph map or plotting of items and students, as well as infit and outfit scattergrams for items and students.

After each run, misfitting items are eliminated, including questions that have been missed by everyone or answered correctly by everyone; student outliers are likewise eliminated. The data are again Rasch analyzed and the winnowing process continued.

Only after this sifting and sorting can the final calibrations and linking be carried out. With every grammar item now linked and calibrated on a common scale, it is possible to create the item bank with questions sorted sequentially from easiest (e.g. a - 3.79 logits) to most difficult (e.g. + 3.91 logits). The initial item bank for BYU's ESL CALT consisted of 300 items (180 grammar and 120 reading). Research, incidentally, has demonstrated the effectiveness of small item pools (Reckase 1983, Tung 1986). Tung indicates "it seems that an item pool not much larger than 100 items of good quality is satisfactory for CAT" (p. 25).

A computer algorithm must next be employed to tailor the test for each student. Algorithms include mathematical versions, notably maximum likelihood estimation procedures, and mechanical or pragmatic versions. The latter was selected because of its speed, economy, practicality, and availability.

The adapted version used with the BYU-ESL CALT divides the 180 grammar items into 22 levels and the 120 reading items into 15 levels. The student begins at approximately the mid-difficulty point. The first five attempts are probes: On the grammar test, students move up four levels for each correct response and down two levels for each incorrect responses (Figure 13). This enables the student to reach the top of the scale within the initial branching routine even if one item is missed. The five reading probes function similarly, except that the student moves up just three levels on correct responses and down two on

incorrect answers. Following the probe routine, students are moved up and down only one level at a time for correct and incorrect responses.

The student is automatically terminated from CALT when four correct responses are recorded at any one level or when five incorrect responses are recorded at any two adjacent levels. At each level, item selection is random and nonrecursive. Placement is at the level where four correct responses have been made or just below the level where a majority of the five incorrect responses have occurred. It is possible to terminate (or "top out" on) either test after completing only seven items, provided all seven initial responses were answered correctly (while nine consecutive incorrect items would be required to "bottom out" on the grammar--eight on the reading).

The present delivery system is the IBM pc with only 64K of memory. In addition, the routine is being adapted for delivery on the Apple II. The present BYU-ESL CALT has been developed for both color and green-screen monitors. While the present test delivers subtests sequentially, it is possible to combine subtest items such as grammar, vocabulary and reading (Larson 1986). A second disk drive is required for a hard copy printout of results. All students using a given microcomputer can have their test results recorded on the same disk.

The present printout includes the student's name, ID number, date, subtest label, the time each subtest was begun and completed, the item number of each test question, its level, which

option was selected, an indication of whether or not the response was right or wrong, the criterion used for level placement (e.g. "There were four correct items at level 16"), and finally the level placement.

Norming is being completed following three administrations of the CALT (in December 1985, and in January and February 1986). The smaller January 1986 administration to 42 incoming students served as a valuable test run since these newcomers to the United States had never had any CAI instruction let alone a computer administered test. The only special preparations made for this group included the placing of color tabs on certain keys (A, B, C, D, the two arrow keys, and "return") and having a bilingual Spanish speaker present as well as a bilingual Japanese speaker.

The 17 male and 25 female students included 17 native Spanish speakers, 17 Japanese, and 8 other. Chi square analysis disclosed that while the Japanese had had more computer exposure than had any of the others, they were significantly more nervous about taking the CALT test than were the Spanish speakers and were less positive about matters ranging from the length of the test to the clarity of the test on the monitor. Nevertheless, after the exam had concluded, 81 percent of the entire group regarded taking the CALT as a good experience--better than comparable ESL/EFL paper and pencil tests.

While the time allocated for taking the conventional grammar test is 25 minutes, the average time to complete the CALT grammar

was 16.9 minutes with a standard deviation of 7.3. Time spent on CALT ranged from a surprisingly brief 3:17 to 32:44. On the conventional reading exam, students are allowed 40 minutes to finish the test. The mean was 27.2 on the CALT reading with a standard deviation of 11.3. The time ranged from just 7:18 to almost an hour--53:3.

While there are 60 items on the conventional grammar test, the average number of CALT grammar items needed to place the student was 20.4 (the standard deviation was 6.8 and the range 22). The number of items on the conventional reading test was 60 on the low battery, 45 on the high. The mean number attempted on the CALT was 22.8, with a standard deviation of 6.6 and a range of 19.

Preliminary results of the CALT administrations indicate that rapid, accurate placement is facilitated together with a substantially reduced testing time for most students. Subsequent research will attempt to verify earlier findings (McBride and Martin 1983, 230-232) that high reliability characterizes the computerized adaptive language test, despite the substantial reduction in the number of items presented.

Microcomputer programs for the one-parameter Rasch model hold out the promise not only of identifying test bias and inappropriate responses on a test but also the means of improving the accuracy and efficiency of language testing through tailored or computerized adaptive measurement. In short, computer-based research applications of latent trait analysis suggest the broad

scope of investigation and refinement now available to the profession.

TABLE 1

ANALYSIS OF VARIANCE: BEGINNING VS LOW INTERMEDIATE AND INTERMEDIATE OUTFIT RESULTS BY TEST TYPE

Test Type	DF	MS	F	P
Reading	1	2.049	19.679	.000*
Listening	1	1.184	21.542	.000
Grammar	1	1.332	14.640	.000

* Significance for the reading test is an artifact of the ANOVA design. Although there was virtually no difference between outfit scores for beginners (level 1) and low intermediate (level2), combining the outfit scores for levels 2 and 3 resulted in a statistically significant difference between levels 1 and the combined scores for levels 2 and 3.

FIGURE 1

ESL Low Battery Listening Test

Map of Student Ability Compared to Item Difficulty

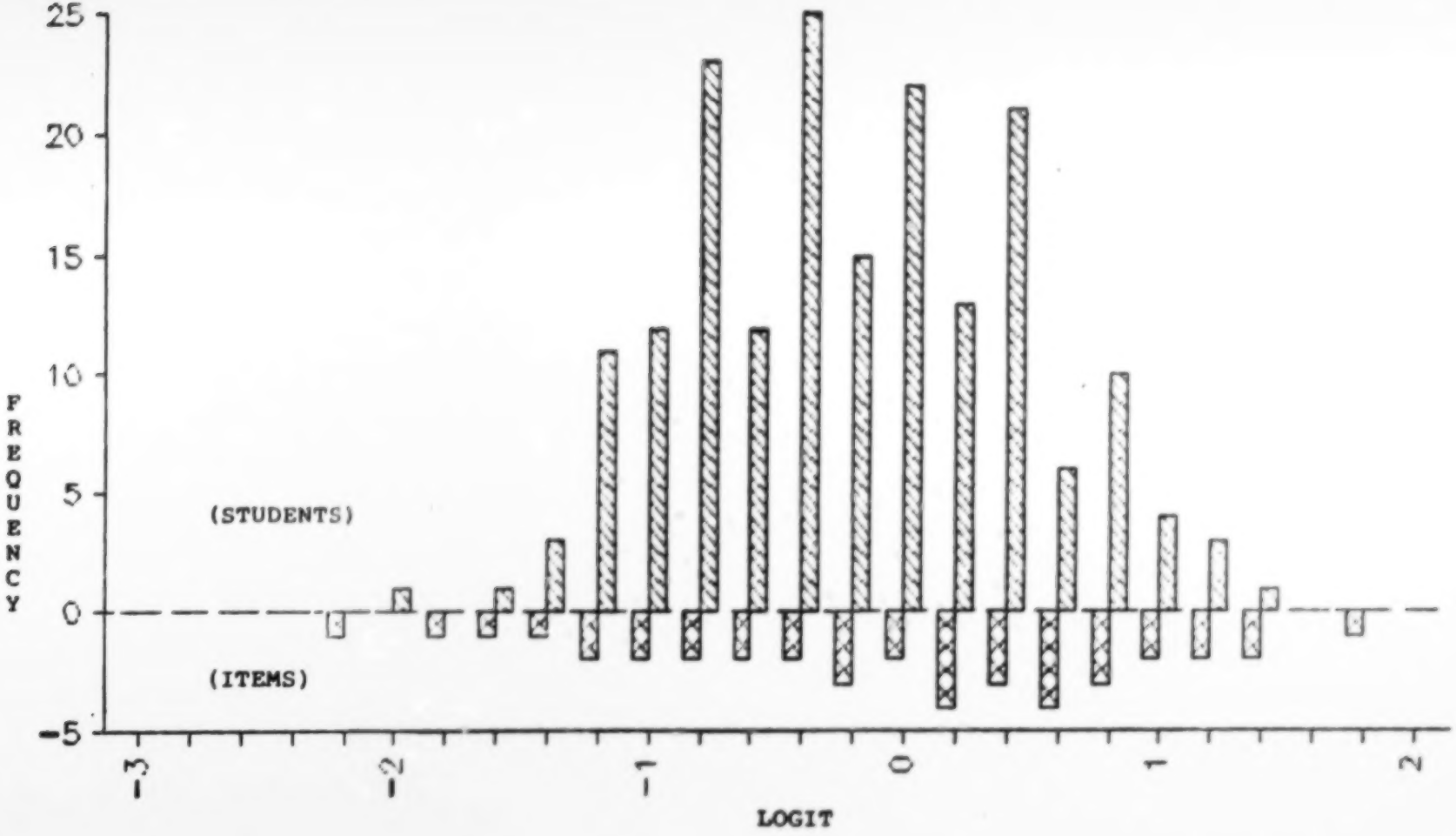


FIGURE 2

Spanish Cognate Vocabulary Exercise

Outfit by Logit for Students

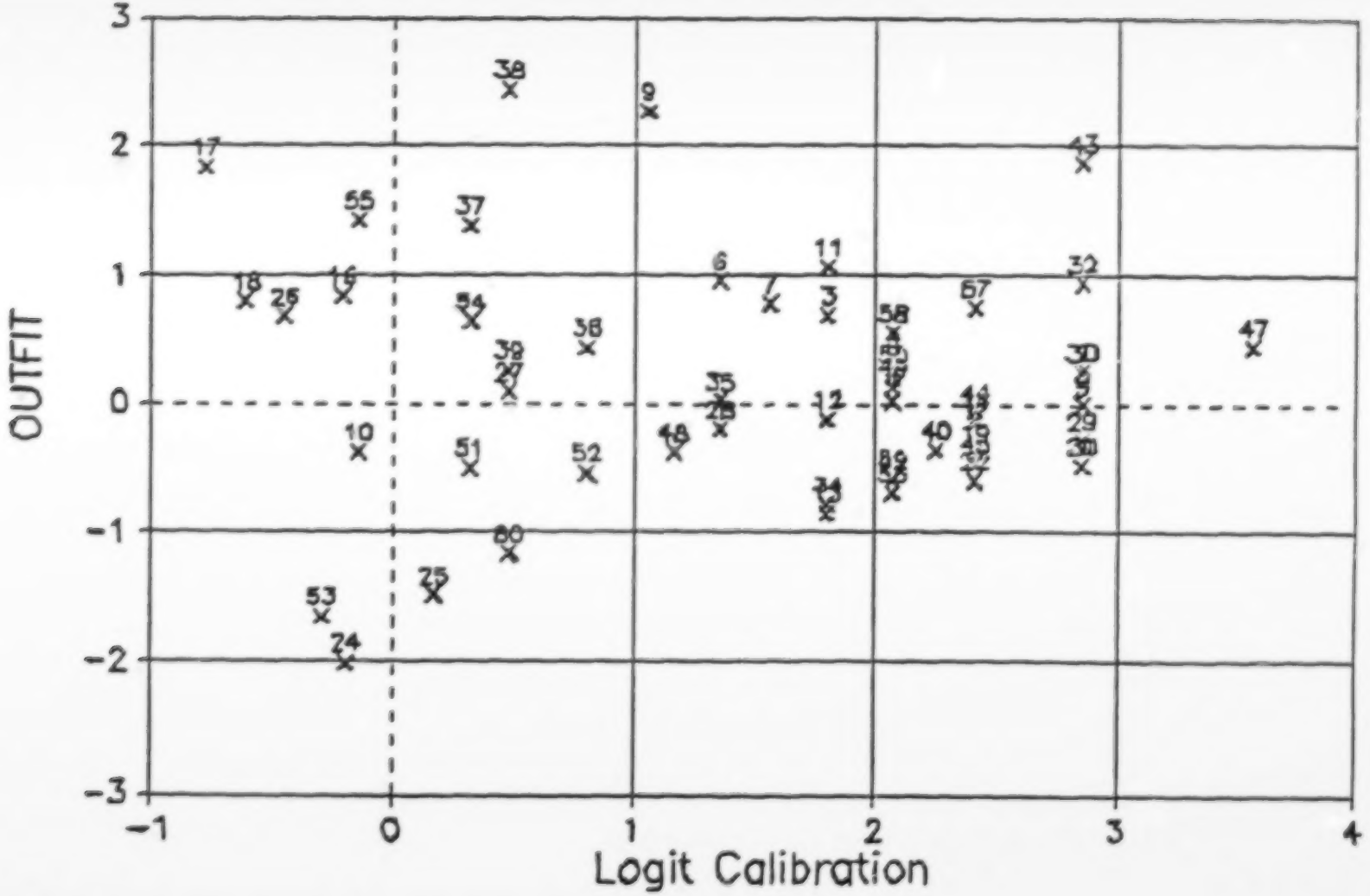
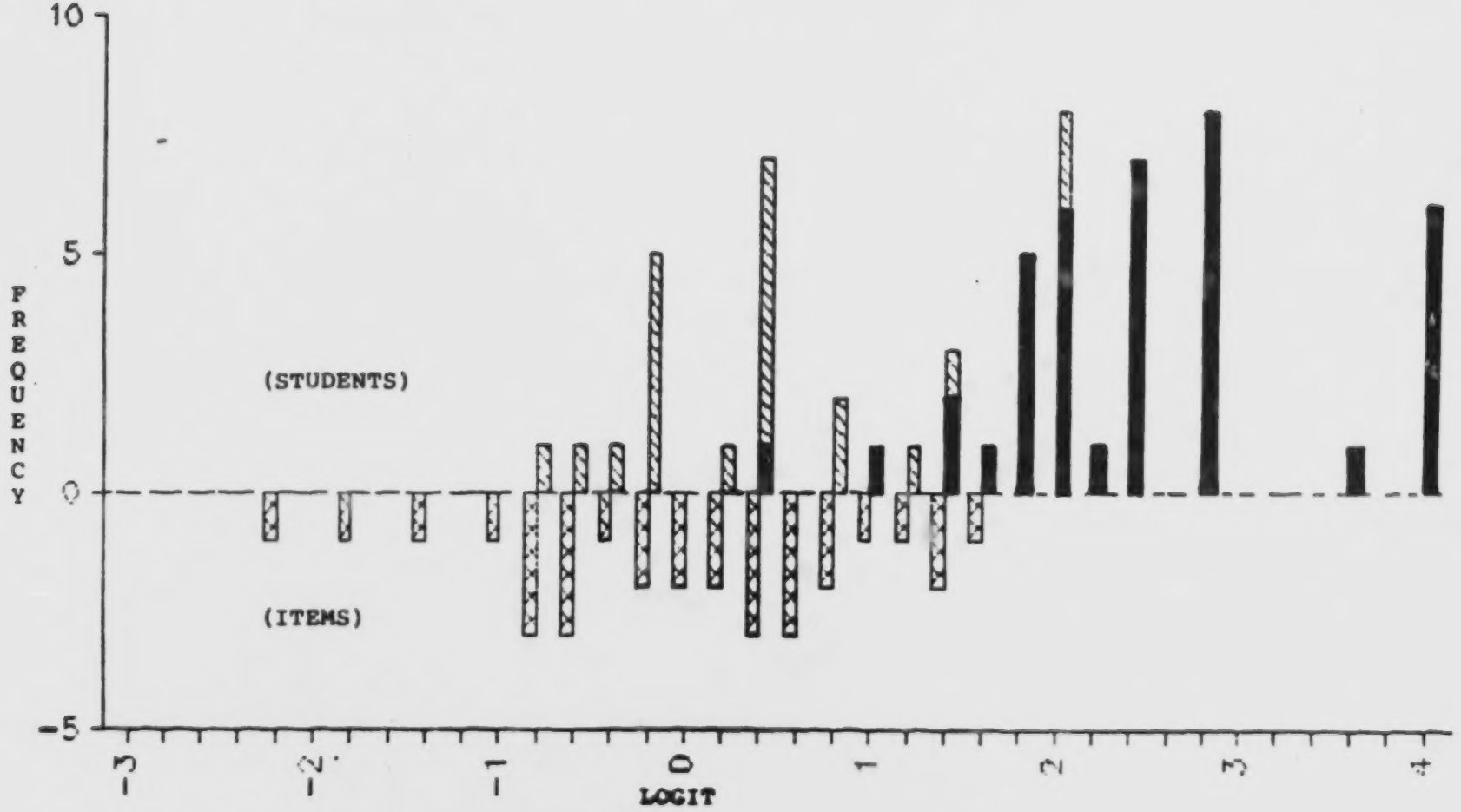


FIGURE 3

Spanish Cognate Vocabulary Exercise

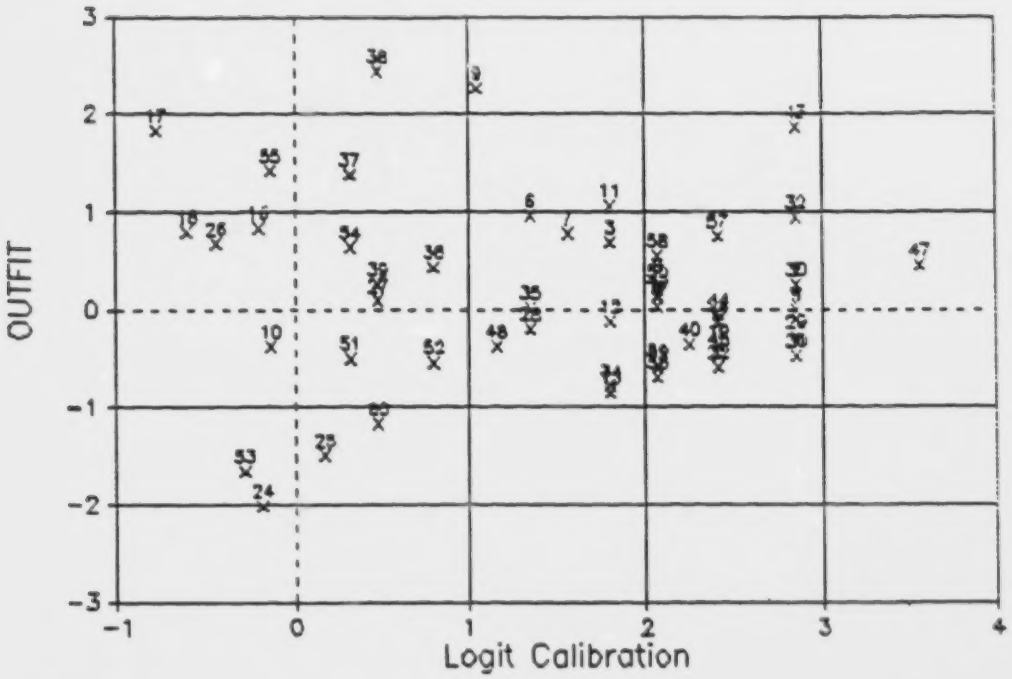
Map of Student Ability Compared to Item Difficulty



NOTE: The solid lines represent students with Spanish-speaking background.

FIGURE 4

Spanish Cognate Vocabulary Exercise
Outfit by Logit for Students



73 percent of outliers are non-Spanish speakers

FIGURE 5

ESL Grammar Test

Outfit by Logit for Items

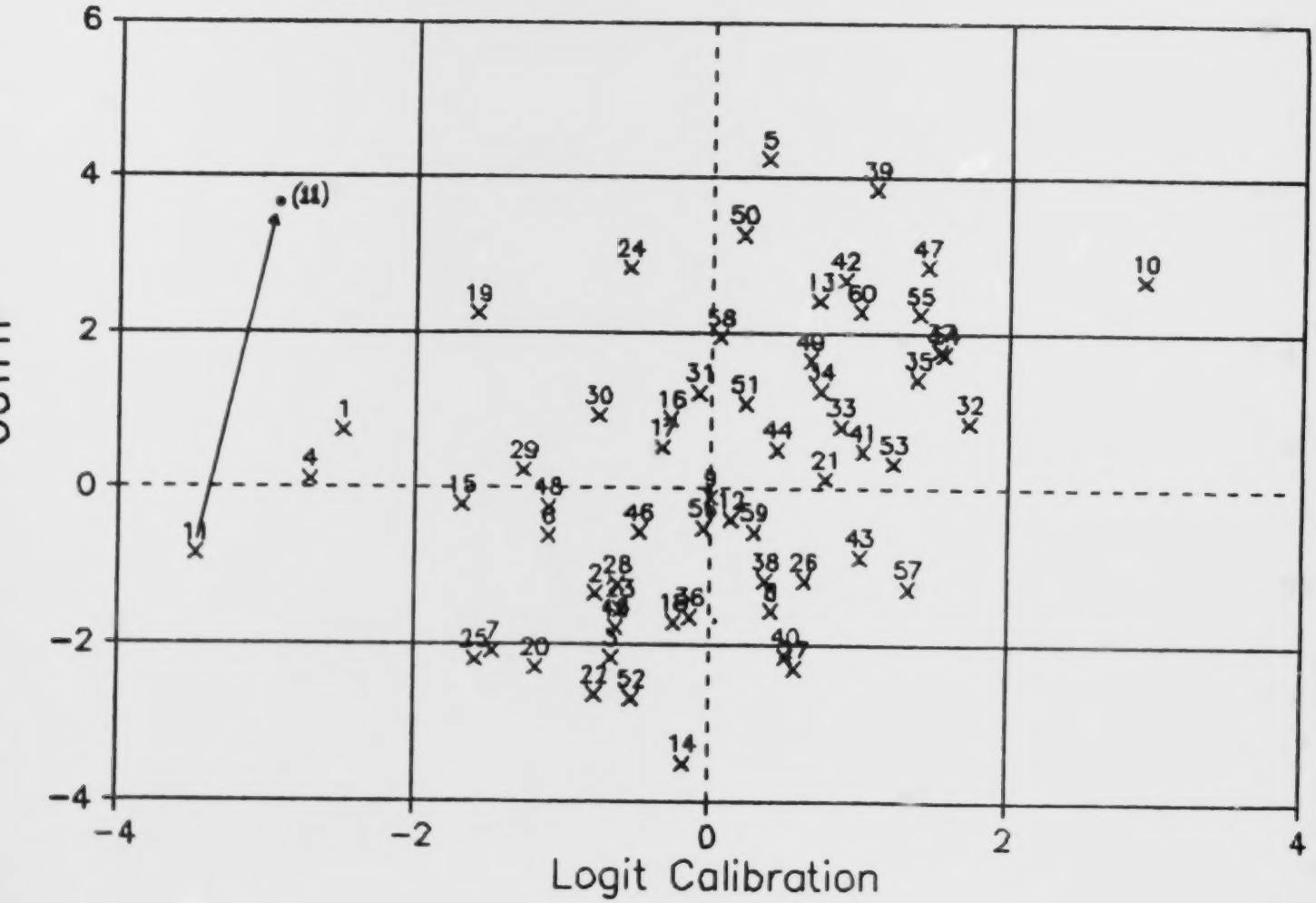


FIGURE 6

ESL Listening Test

Student Distribution by Language

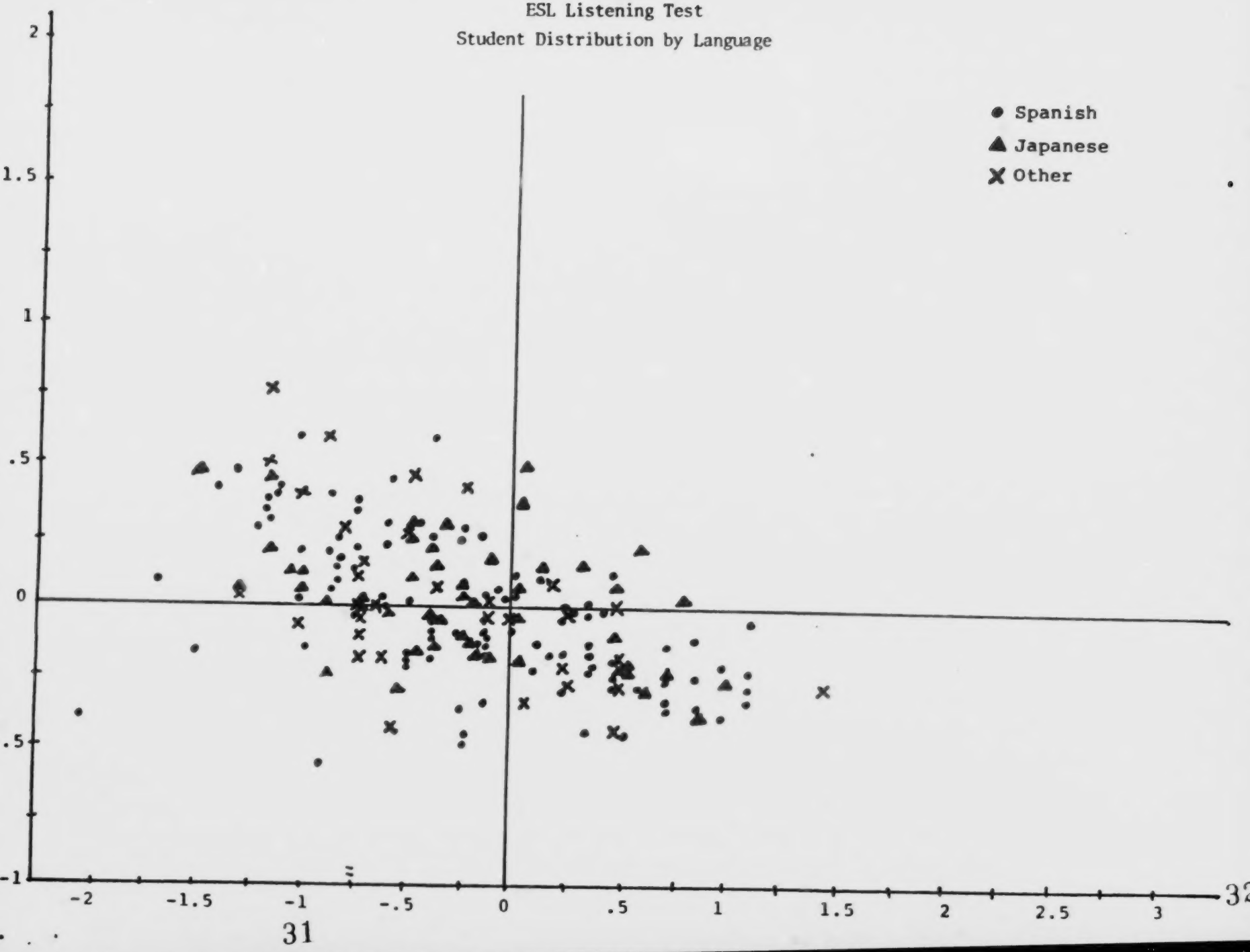


FIGURE 7

ESL Grammar Test

Student Distribution by Language

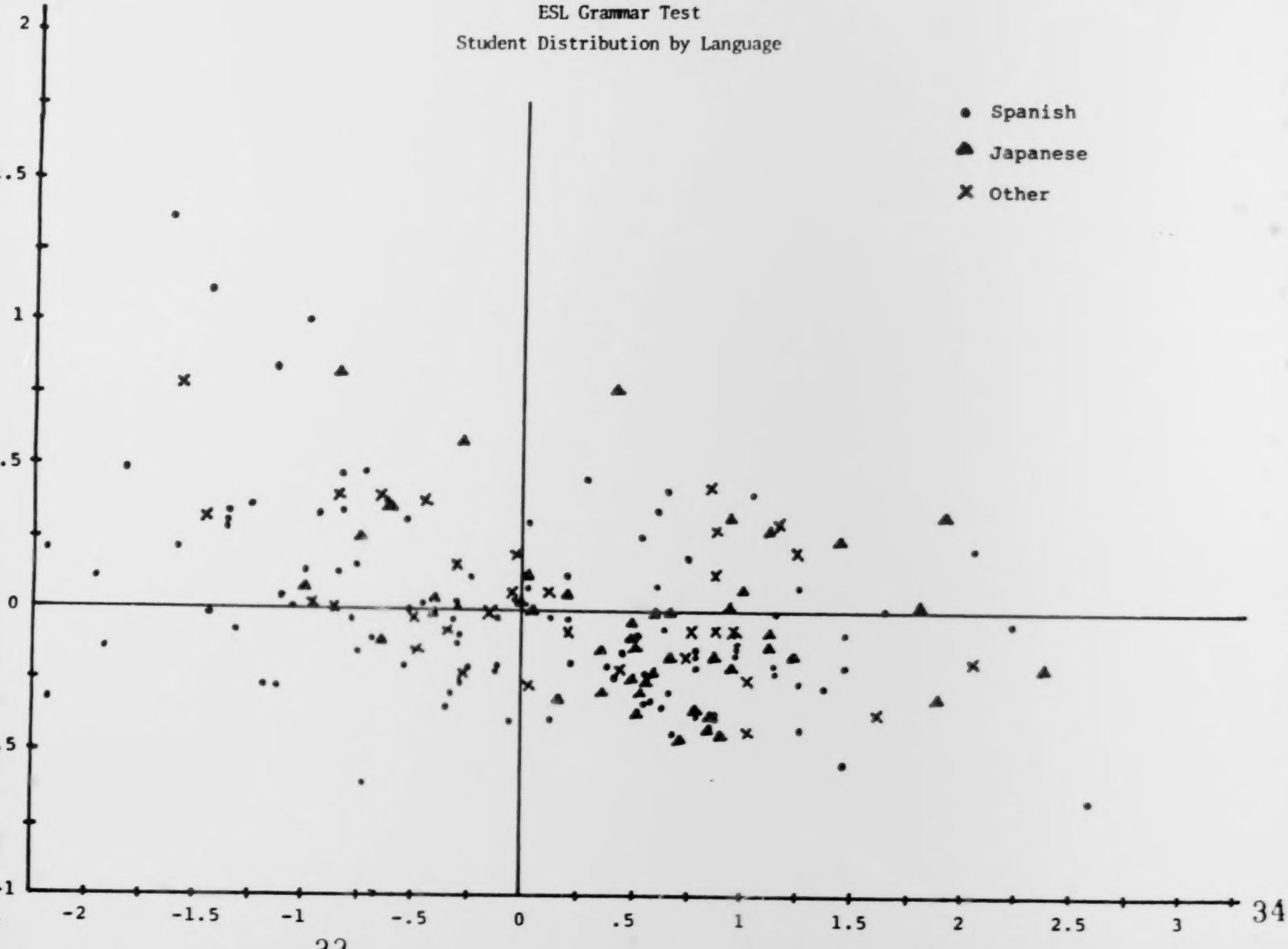
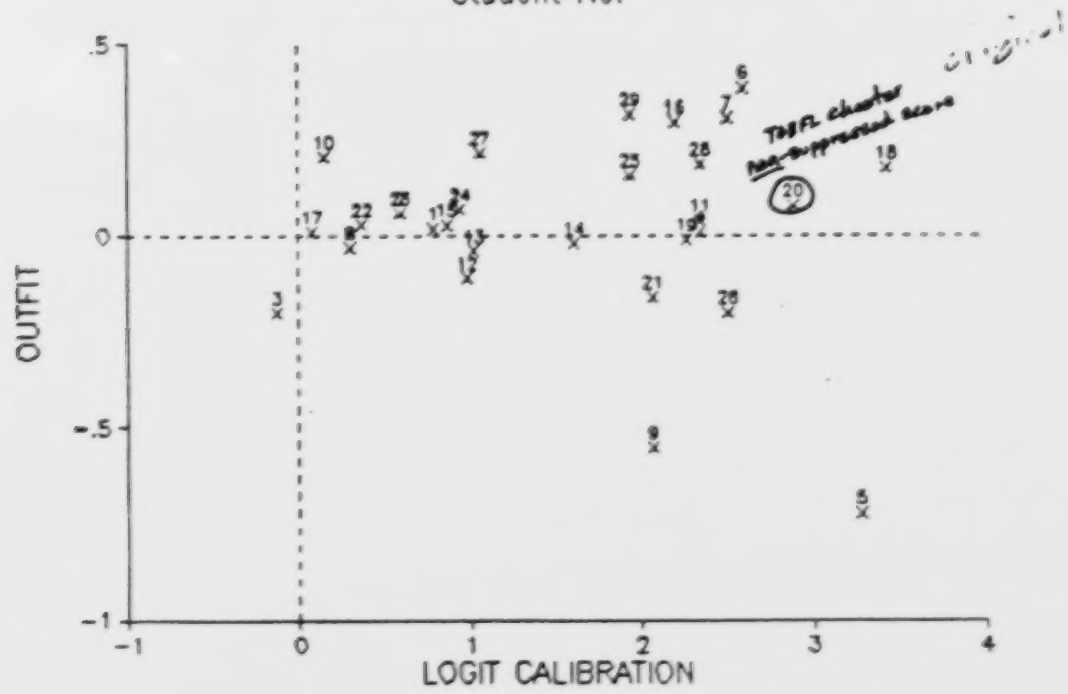


FIGURE 8

TOEFL(7) EXAM DATA ON SUPPRESSED SCORES



TOEFL(15) EXAM DATA FOR SUPPRESSED SCORE

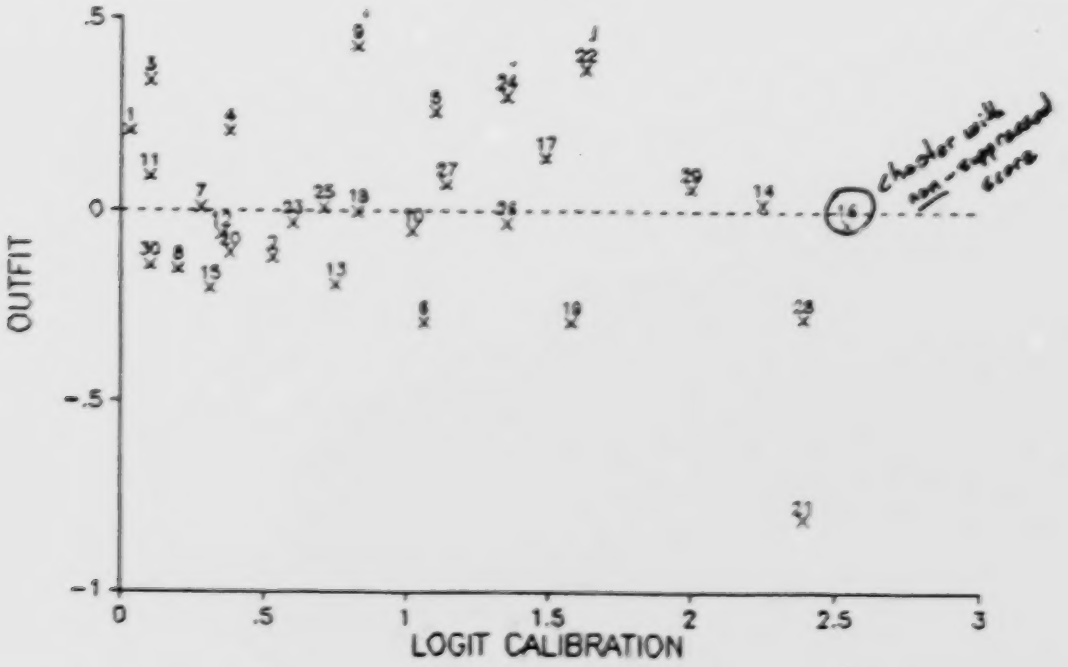
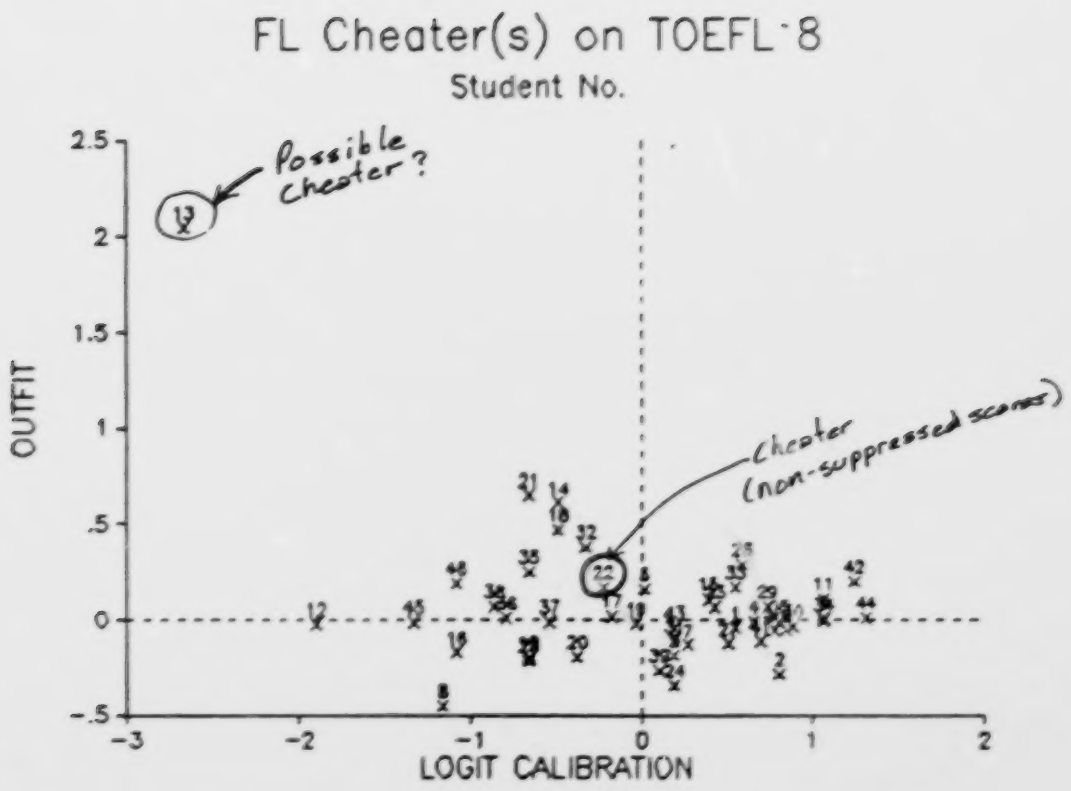


FIGURE 9


FL Cheater(s) on TOEFL 8
Student No.
Possible cheater?
Cheater (non-suppressed scores)
OUTFIT
LOGIT CALIBRATION
2.5
2
1.5
1
.5
0
-.5
-3
-2
-1
0
1
2

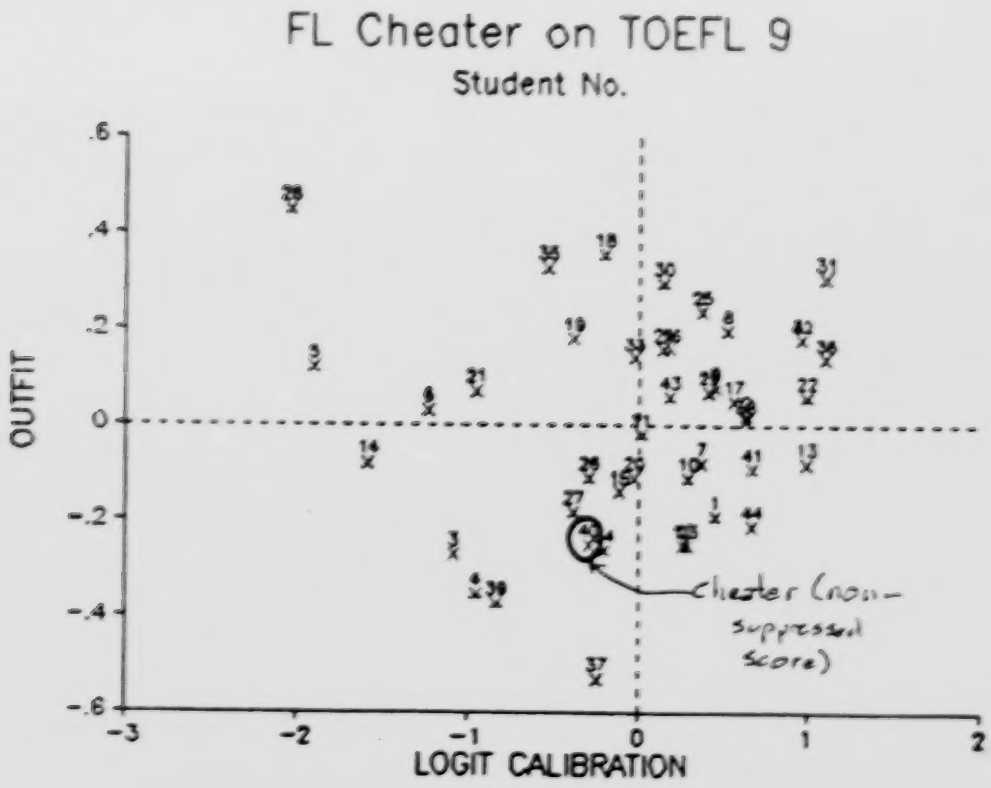

FL Cheater on TOEFL 9
Student No.
Cheater (non-suppressed score)
OUTFIT
LOGIT CALIBRATION
.6
.4
.2
0
-.2
-.4
-.6
-3
-2
-1
0
1
2

FIGURE 10

ESL Grammar Test

Outfit by Logit for Students

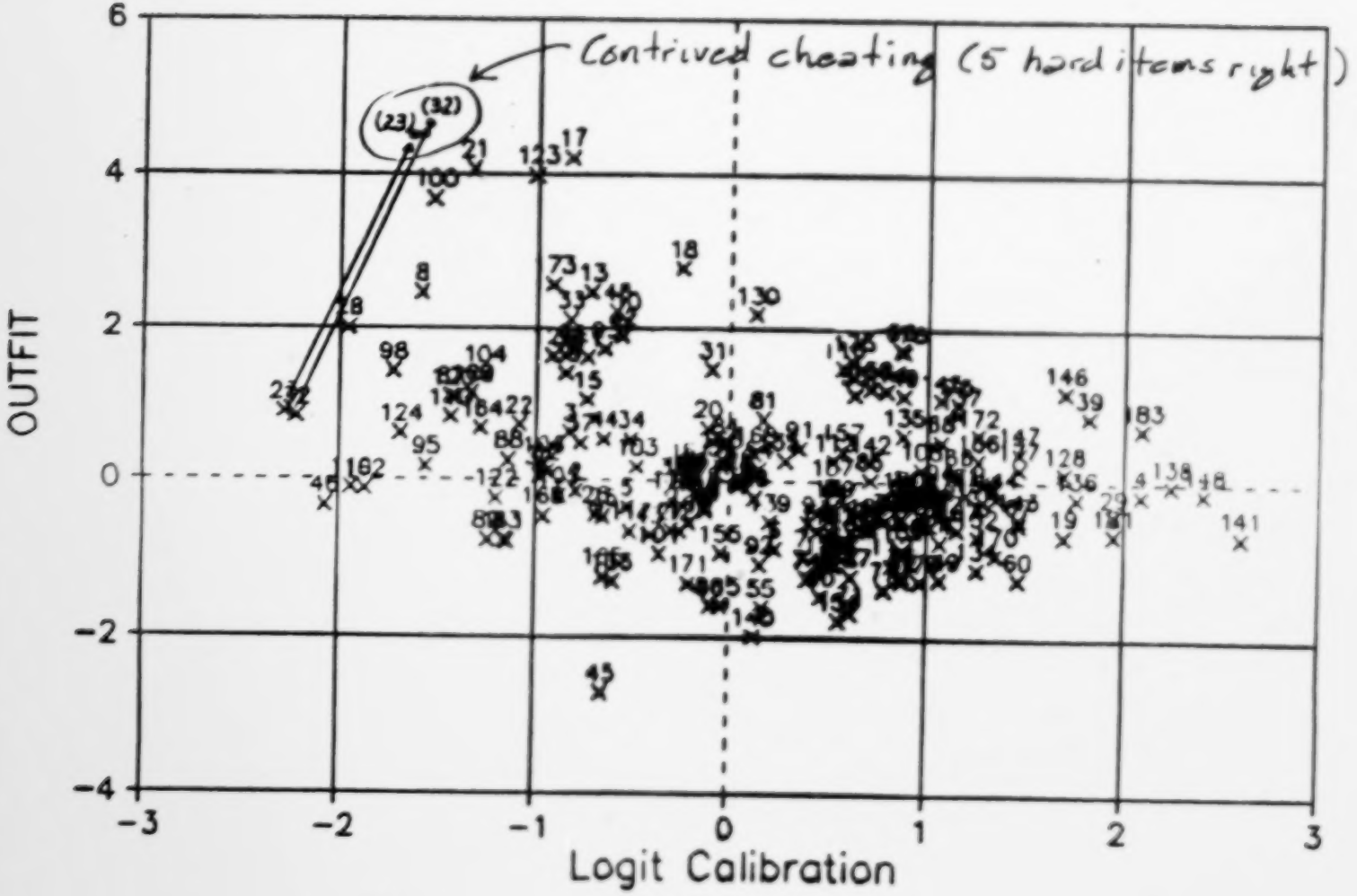
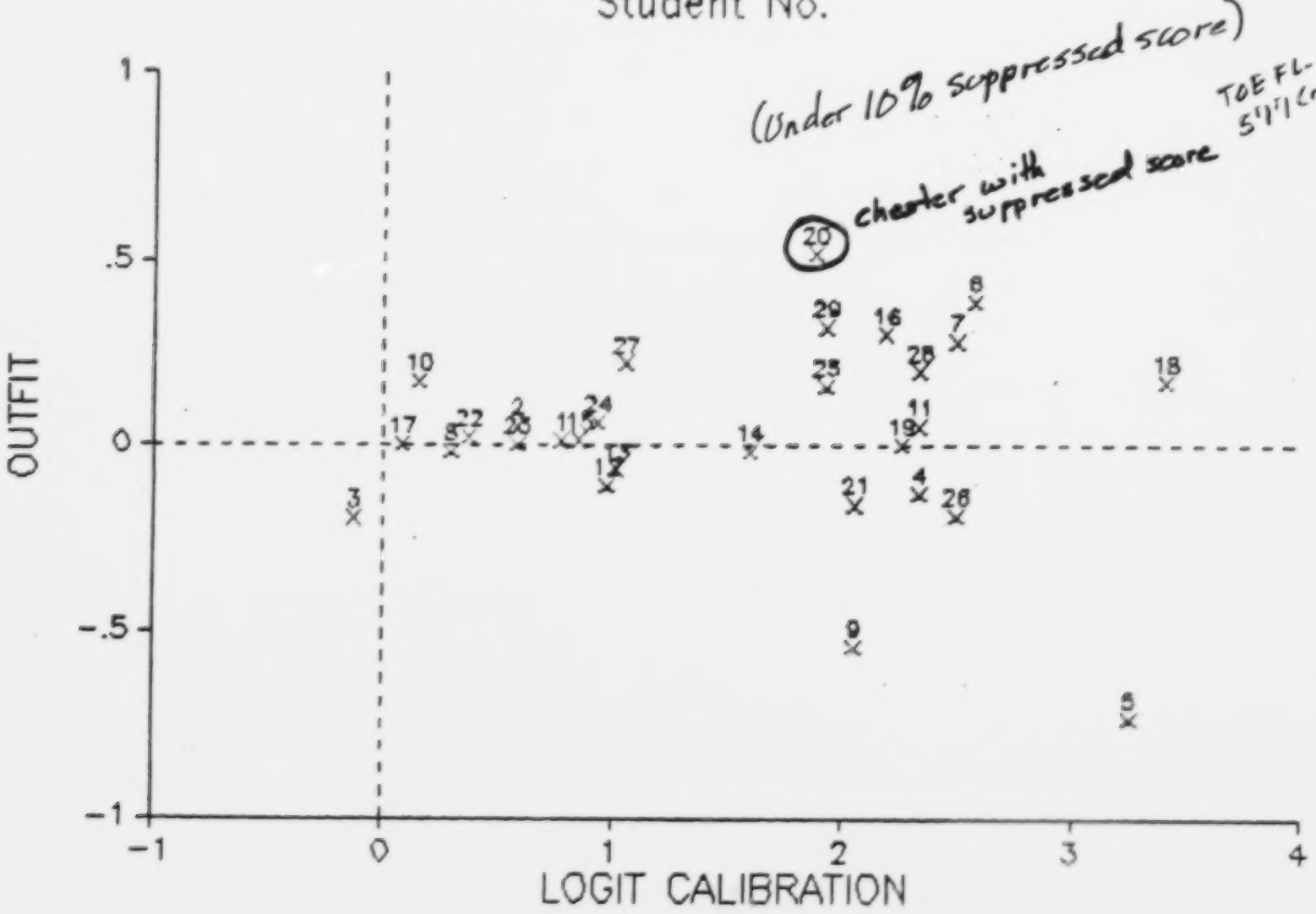


FIGURE 11

CONTRIVED FL TOEFL7 'CHEATER': SUPPRESSED



TOEFL (7)

CONTRIVED NL TOEFL CHEATER: SUPPRESSED

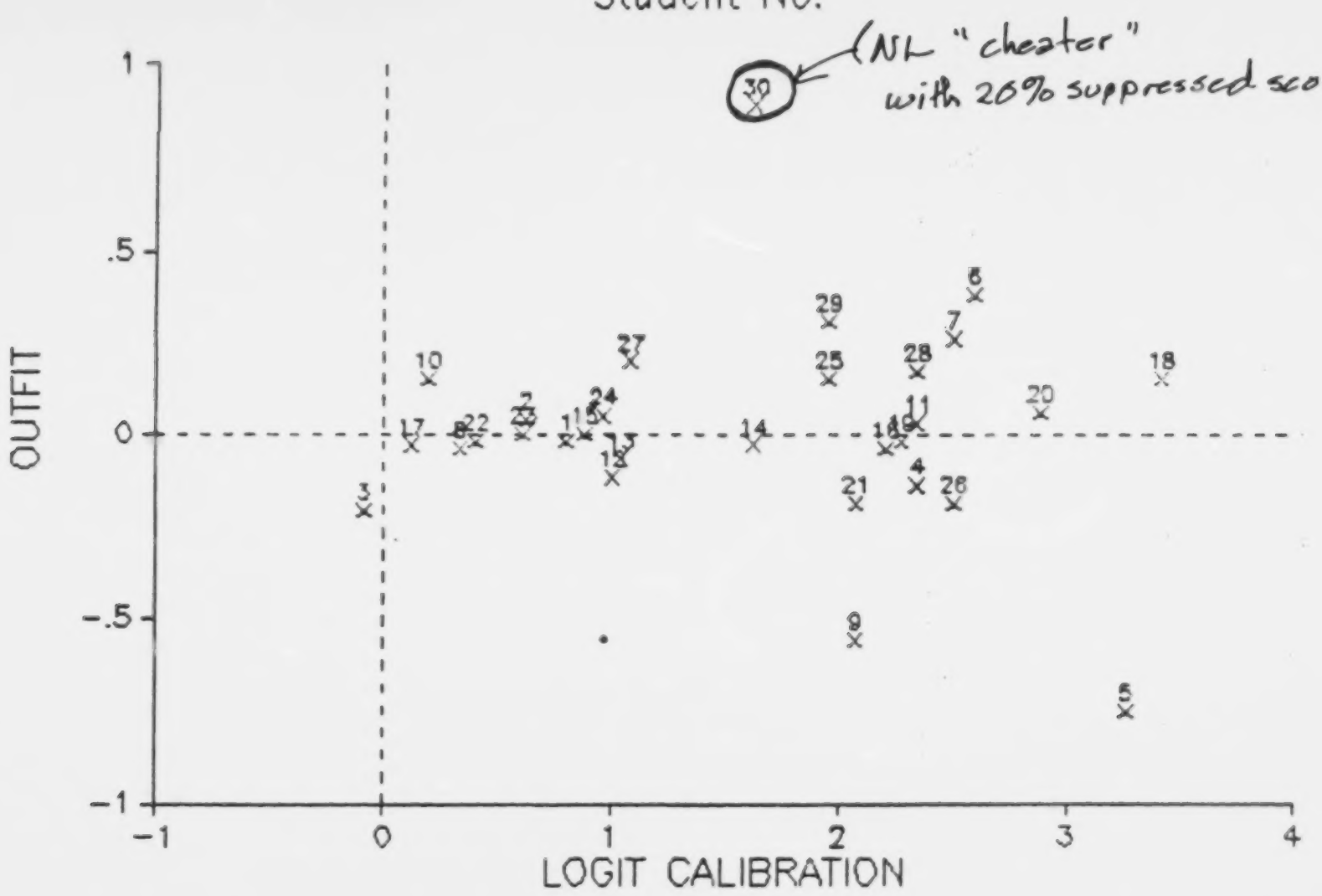


FIGURE 12

REFERENCES CITED

Alfred, Bertha. 1986. Personal communication: test development at the American University in Cairo (February 28).

Briere, Eugene J. 1973. "Cross-cultural Biases in Language Testing." In Focus on the Learner: Pragmatic Perspectives for the Language Teacher, John W. Oller, Jr. and Jack C. Richards (Eds.), 214-228. Rowley, Massachusetts: Newbury House Publishers, Inc.

Canale, Michael. 1986. "The Promise and Threat of Computerized Adaptive Assessment of Reading Comprehension." In Technology and Language Testing, Charles W. Stansfield (Ed.), 29-44. Washington, D.C.: TESOL.

Davidson, Fred and Grant Henning. 1985. "A Self-rating Scale of English Difficulty: Rasch Scalar Analysis." Language Testing 2:2 (December), 164-179.

Englehard, George, Jr. 1980. "An Introduction to Rasch Measurement and Its Application to Test Equating in the Comprehensive Assessment Program." ED 194 636.

Farhady, Hossein. 1979. "Test Bias in Language Placement Examinations." In On TESOL '79, Carlos A. Yorio, Kyle Perkins, and Jacquelyn Schachter (Eds.), 162-170. Washington, D.C.: TESOL.

Goldstein, H. 1979. "Consequences of Using the Rasch Model for Educational Measurement." British Educational Research Journal. 5:211-220.

Henning, Grant and Fred Davidson. 1986. "Scalar Analysis of Composition Ratings." Paper presented at the Eighth Language Testing Research Colloquium, Monterey, Ca.; February 28.

Henning, Grant, Thom Hudson and Jean Turner. 1985. "Item Response Theory and the Assumption of Unidimensionality for Language Tests." Language Testing 2:2 (December), 141-154.

Houts, P. L. (Ed.) 1977. The Myth of Measurability. New York: Hart.

Hulin, Charles L., Fritz Drasgow, Charles K. Parsons. 1983. Item Response Theory: Applications to Psychological Measurement. The Dorsey Professional Series. Homewood, Illinois: Dow Jones-Irwin.

Larson, Jerry W. 1986. Demonstration of a computerized-adaptive Spanish test. Eighth Annual Language Testing Research Colloquium, Monterey, California; February 28.

Larson, Jerry W. and Harold S. Madsen. 1985. "Computerized Adaptive Language Testing: Moving Beyond Computer-Assisted Testing." CALICO Journal 3(2): 32-36, 43.

Levine, Michael V. and Fritz Drasgow. 1982. "Appropriateness Measurement: Review, Critique and Validating Studies." British Journal of Mathematical and Statistical Psychology 35:42-56.

Levine, Michael V. and D. F. Rubin. 1979. "Measuring the Appropriateness of Multiple-choice Test Scores." Journal of Educational Statistics. 4:269-290.

Lord, Frederic M. 1983. "Small N Justifies Rasch Model." In New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing, David J. Weiss (Ed.), 51-61. New York: Academic Press.

Madsen, Harold S. 1982. "Determining the Debilitative Impact of Test Anxiety." Language Learning 32 (1): 133-143.

Madsen, Harold S. and Jerry W. Larson. 1986. "Computerized Rasch Analysis of Item Bias in ESL Tests." In Technology and Language Testing, Charles W. Stansfield (Ed.), 47-67. Washington, D.C.: TESOL.

McBride, James R. and John T. Martin. 1983. "Reliability and Validity of Adaptive Ability Tests in a Military Setting." In New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing, David J. Weiss (Ed.), 223-236. New York: Academic Press.

Mohan, Bernard. 1979. "Cultural Bias in Reading Comprehension Tests." In On TESOL '79, Carlos A. Yorio, Kyle Perkins, and Jacquelyn Schachter (Eds.), 171-177. Washington, D.C.: TESOL.

Nakano, Patricia J. 1977. "Educational Implications of the Lau v Nichols Decision." In Viewpoints on English as a Second Language. Marina K. Burt, Heidi C. Dulay, and Mary Finocchiaro (Eds.), 219-234. New York: Regents Publishing Co.

Oller, John W., Jr. and Kyle Perkins. 1978. Language in Education: Testing the Tests. Rowley, Massachusetts: Newbury House Publishers, Inc.

Rasch, Georg. 1966. "An Item Analysis which Takes Individual Differences into Account." British Journal of Mathematical and Statistical Psychology. 19:49-57.

Reckase, Mark D. 1983. "A Procedure for Decision Making Using Tailored Testing." In New Horizons in Testing: Latent

Trait Test Theory and Computerized Adaptive Testing, David J. Weiss (Ed.), 237-255.

Rubin, Lois S. and David E. W. Mott. 1983. "Comparison of Three Different Item Banking Methods for Longitudinal Test Equating." (April) ED 228 315.

Tung, Peter. 1986. "Computerized Adaptive Testing: Implications for Language Test Developers." In Technology and Language Testing, Charles W. Stansfield (Ed.), 11-28. Washington, D.C.: TESOL.

Weiss, David J. (Ed.) 1983. New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. New York: Academic Press.

Woods, Anthony and Rosemary Baker. 1985. "Item Response Theory." Language Testing 2:2 (December), 119-140.

Wright, Benjamin D. and Michael Linacre. 1985. Microscale Manual. Westport, Ct.: Mediax Interactive Technologies.

Wright, Benjamin D. and Geofferey N. Masters. 1982. Rating Scale Analysis: Rasch Measurement. Chicago: Mesa Press.

Wright, Benjamin D. and R. J. Mead. 1976. "BICAL: Calibrating Items with the Rasch Model." Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago.

Wright, Benjamin D. and Mark H. Stone. 1979. Best Test Design. Chicago: Mesa Press.